



ASHESI UNIVERSITY COLLEGE

USING MACHINE LEARNING TO IMPROVE PUBLIC HEALTH SURVEILLANCE IN LOW-INCOME COUNTRIES.

Disease Forecasting and Disease Outbreak Detection using Open-source technology

NUPIC by NUMENTA.

Undergraduate Thesis

B.Sc. Computer Science

Salifu Mutaru

2016

ASHESI UNIVERSITY COLLEGE

Using Machine Learning to improve public health surveillance in low-income countries.

Undergraduate Thesis

Undergraduate Thesis submitted to the Department of Computer Science,
Ashesi University College in partial fulfilment of the requirements for the
award of Bachelor of Science degree in Computer Science.

Salifu Mutaru

April 2016

DECLARATION

I hereby declare that this Undergraduate Thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

.....

Candidate's Name:

Mutaru Salifu

Date:

3rd May, 2016

I hereby declare that preparation and presentation of this Undergraduate Thesis were supervised in accordance with the guidelines on supervision of Undergraduate Thesis laid down by Ashesi University College.

Supervisor's Signature:

.....

Supervisor's Name:

Mr. Kwadwo Gyamfi Osafo-Mafo

Date:

.....

Acknowledgement

This research would not have been successful without the support of Mr. Scott Purdy of Numenta. Mr. Scott explained how to access some functionality of Numenta's implementation of Hierarchical Temporal Memory Algorithms. I have also benefited greatly from the supervision of Mr. Kwadwo Gyamfi Osafo-Mafo—his questions inspired some of the experiments conducted in this research. Mr. Aelaf Dalfa helped me get the data I need for this research, and his questions laid out the plans for the future of this research. I deeply appreciate Grameen Foundation Ghana and the Ashesi-mHealth program for providing me the data I needed for this research.

There are no words I can use to express my gratitude towards my family for understanding the distance I needed to complete this work, and of course the prayers they sent me each day throughout this period. I also thank my godfather Mr. Samuel Aryee of the Accra Metropolitan Assembly for all the support he has given me throughout my life in college. Most important of all, I send sincere gratitude to the MasterCard Foundation for funding my education at Ashesi University College. Without MasterCard Foundation I would not have dreamt of this research in the first place.

Abstract

Public health surveillance is a challenge across the globe. Unexpected disease outbreaks occur even though we have vast knowledge in the field of medicine. Also not being able to forecast the next set of diseases to be reported keeps our medical practitioners less prepared. This problem is heightened in low-income countries that cannot afford the infrastructure that comes with ubiquitous computing methods including Machine Learning on a national scale. There is also a notion that such computing approaches by huge companies such as Google and IBM are too expensive to implement. This research seeks to investigate if Machine Learning can be used in low-income countries to provide an effective public health surveillance regardless challenges in funding and availability of ubiquitous infrastructure. Disease Outbreak Detection and Disease Forecasting are the two branches of public health surveillance considered in this research.

A prototype application called MATE is built based on Numenta's implementation of Hierarchical Temporal Memory (HTM) Algorithms. Experiments were conducted with MATE proved 75% to 95% accuracy for forecasting next 20 diseases, and an instant recognition of anomalies for Disease outbreak detection.

Keywords: Biosurveillance, Machine Learning, Hierarchical Temporal Memory

Table of Contents

DECLARATION	iii
Acknowledgement.....	iv
Abstract.....	v
List of Figures.....	viii
List of Tables	ix
Chapter 1 : Introduction	1
Chapter 2 Literature Review	6
2.1 Discussion	10
2.1.1 Limitations of these Theses	11
Chapter 3 Method	17
3.1 Implementation	17
3.1.1 How MATE works.....	19
3.1.2 How Prediction works.....	20
3.1.3 How Anomalies are detected by MATE	23
3.1.3 Raw Anomaly Score	23
3.2 Dataset.....	23
3.2.1 Data from Numenta.....	23
3.2.2 Data from Ashesi-mHealth Project	24
3.2.3 Inpatients reported cases from Grameen Foundation	24
3.3 Experiments	24
3.3.1 Experiment 1(a)	25

3.3.2 Experiment 1(b)	27
3.3.3 Experiment 2(a)	31
3.3.4 Experiment 2(b)	35
Chapter 4 : Conclusion.....	38
References	39

List of Figures

Figure 3-1 Architecture of HTM.....	17
Figure 3-2 Architecture of MATE	19
Figure 3-3 A layer of cells in HTM for first data point	20
Figure 3-4 A layer of cells in HTM when analyzing familiar data points	21
Figure 3-5 ARI actual vs. ARI predicted in skewed data, community4	25
Figure 3-6 Malaria actual vs. Malaria predicted in skewed data, community4	25
Figure 3-7 Malaria actual vs. Malaria predicted in skewed data, community17	26
Figure 3-8 ARI actual vs. ARI predicted in skewed data, community17	26
Figure 3-9 Many anomalies found in new dataset	32
Figure 3-10 Fewer anomalies found as MATE learns	33
Figure 3-11 No anomalies are found after MATE learns enough patterns	33
Figure 3-12 Anomalies found on their first occurrence	34
Figure 3-13 Anomalies discovered again not as values but as patterns	35
Figure 3-14 MATE does not find earlier anomalous data as anomalous after learning patterns of the anomaly.	36
Figure 3-15 MATE begins to predict earlier detected anomaly after learning enough patterns in the occurrence of the anomaly.....	36

List of Tables

Table 2-1 Communities and their data size.....	14
Table 3-1 Good predictions with increasing accuracy in evenly distributed data	28
Table 3-2 Accuracy of MATE data increases	31

Chapter 1 : Introduction

Public health surveillance is a global challenge today, but the challenge is greater in low-income countries that cannot afford the infrastructure that comes with ubiquitous computing. This research investigates possible techniques of Machine Learning that can be used in low economic zones to automate Disease Outbreak Detection and Forecasting of diseases over a period of time. Disease outbreak detection depends on the effectiveness of an algorithm to discover anomalies in health data. Anomalies signal a sign of an outbreak which can be further studied to confirm if it is simply noise in data. Forecasting on the other hand is based on predictive algorithms that can predict over n steps. Thus, instead of predicting just the next occurrence, forecasting algorithms predict more than one next occurrences.

The problem of Disease Outbreak Detection and Forecasting of diseases has been tackled using different methods that largely depend on supervised learning algorithms. This research studies these approaches to investigate their feasibility in low-income areas.

Cooper, Dash, Levander, Wong, Hogan, Wagner (2004) used Causal Bayesian Networks to model patients of a population. Each patient's attributes was modeled in a 14-node sub network. Cooper et al decided to use a single node for patients with same attributes. Based on the current health status of a patient, her node in the network signals a probability of inferring a disease. If more other nodes within the network signal the same disease with a probability greater than an agreed threshold, then the Bayesian network Biosurveillance system signals an expected Disease outbreak.

The research of Cooper et al is an improvement of their work in 2003 entitled Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. Wong, Moore, Cooper, Wagner (2003) proposed that, instead of determining outbreaks by comparing

distribution of recent data against a baseline distribution, baseline distribution trends must be attributed with a conditioning that gives a probabilistic feature of each possible outbreak based on current conditions such as weather and seasons. Thus, for each discovered outbreak, Cooper et al agreed the probability of the discovery being a true outbreak must inform the final conclusion. Each rule discovered in data is scored with a conditional value p based on its relation to actual baseline distribution. Wong et al proved that with scored probabilities for each rule, their system called WSARE 3.0 is able to discover the most significant possible outbreak for each day based on baseline distribution on that day. With scored probabilities for each rule, false positive count was reduced, and WSARE 3.0 was able to detect outbreaks in simulated data with the almost the earliest possible detection time while keeping a low positive count.

Kaustav, Schneider, Neill (2008) believe that anomalies are generated by underlying processes which are generated by particular subsets of data. Searching a whole record of data for an anomaly is therefore not effective. Kaustav et al created separate local anomaly detectors specific to the subsets of each record of data. When a local anomaly detector discovers an anomalous pattern, same subsets of other rows of data are investigated if they have similar anomalous pattern. The approach of Kaustav et al depends on Bayesian Network Likelihood and Conditional Anomaly Detection.

Social media platforms such as Facebook and Twitter are continuously growing into richer sources of data for research. Yang, Cui, Hu, Zhu, Yang (2014) used social media analysis for disease surveillance in Beijing. The team of researchers collected social media posts from a Chinese micro-blog. The collected posts were classified using two methods, K-Means for unsupervised learning and Support Vector Machine (SVM) for supervised learning. Line charts were drawn for the results of the clusters that related to Influenza. The

results of the line charts were compared to weekly reports of the National Influenza Center. Yang et al found SVM (supervised method) to be more efficient since the data available was not enough for the unsupervised K-Means approach to produce equally good predictions. The interesting aspect of this research is that it is the first to analyze Chinese text from a social media network for Biosurveillance—the results could both detect disease outbreaks and forecast diseases over a period of time.

Kaundal, Kapoor, Raghav (2006) propose that since neural networks and multiple regression are unable to predict value of unknown data points and also require longer training times, Support Vector Machines is a better approach to forecast diseases of plants, putting into account factors such as weather. Historical data on farm diseases over the year 2000 was gathered and correlated with meteorological data. Weather variables such as rainfall, temperature, relative humidity, sunshine and etc. then helped predict the next set of diseases expected based on the weather conditions that are expected (weather conditions and other conditions that correlate with diseases of people, can be used in another case), and their correlation and causation to past data. Different approaches including; Artificial Neural Networks, Multiple Regression, and Support Vector Machines were tested on the data. From Kaundal et al, SVM produced the most accurate forecast with the shortest learning period.

Last but not least, Rana, Gupta, Phung, Venkatesh (2015) disapproved of using SVM, Naive Bayes or Random Forest as predictive frameworks for clinical interventions. This is because, the mentioned approaches amalgamate input variables into a single rule for all their predictions. Clinical interventions however evolve as new diseases and new approaches to solving diseases, or even new understanding of diseases are discovered. Rana et al rather proposed a predictive framework that separates interventions from patient condition. That way, just like in OOP, changes can be made to the separated interventions

as they evolve so to understand current trends in patient data in order to make the right predictions—thus, an evolving rule for prediction is created instead of a static set of rules.

The papers reviewed here show the amount of research going into Machine Learning in order to automate complex problems in Biosurveillance. All approaches used in the researches reviewed are based on supervised learning models, which is safe in the case of Biosurveillance. Biosurveillance has to be as accurate as possible, using unsupervised learning approaches although have their advantages such as discovering unexpected patterns and functioning without training set, may miss key rules expected to detect specific trends and diseases. Yang et al confirmed this when their text experiment on data from Chinese micro-blog produced more accurate forecast with SVM (supervised learning model) compared to K-Means (unsupervised learning model).

In the case of low-income countries however, it should be merrier to build a Biosurveillance system that is very cost effective in implementation and maintenance. This paper does not in effect advise low-income countries not to invest in research and development in areas such as public health surveillance, but rather motivates such investments by proving how affordable such interventions can be. This research will therefore investigate if unsupervised learning model based on the open-source Hierarchy Temporal Memory (HTM) system by NUMENTA, can be used in such communities to produce good predictions for forecasting diseases and detecting outbreaks.

HTM approach is chosen because NUMENTA have provided an open-source version of the implementation of HTM called NUPIC (Numenta Platform for Intelligent Computing), which makes it easy to implement. NUPIC is also chosen because with the same algorithm, temporal and spatial patterns can be learned, prediction and modeling of data is achieved, patterns are clustered, and a benchmark anomaly detection algorithm is in-

built. Also, NUPIC is a generic Machine Learning system that many researches are undergoing currently to find out what it can be used for, and what its limits are. There is presently no research to find out if HTM algorithms as implemented in nupic can be used in Biosurveillance. HTM algorithms have been used to model servers and other machines to predict energy usage and anomalies that predict an expected breakdown. HTM algorithms by NUMENTA have been used to find anomalies in public traded companies, discover behavior changes of workers that may lead to internal insecurity. This research not only investigates if unsupervised learning models can be employed for effective Biosurveillance in areas with challenges such as funding, availability of training dataset and ubiquitous infrastructure, but also investigates if Hierarchical Temporal Memory algorithms implemented in nupic by Numenta is a feasible unsupervised learning approach for Biosurveillance. This decision is highly motivated by the need to build a single system of interacting algorithms that can achieve most aspect of Biosurveillance, including disease forecasting and disease outbreak detection.

Chapter 2 Literature Review

The current way of administering health-care in low economies such as Ghana depends largely on the efficiency of the doctor and supporting nurses. Obviously health practitioners are not flawless regardless their years of experience. Detecting disease outbreaks, forecasting expected diseases and other forms of public health surveillance require computers to make sense of huge accumulated data. Today, systems such as the IBM Watson for Health have proven even diagnosis can be vastly improved with ubiquitous cognitive computing (“IBM Watson”, 2016). This form of computing is largely achieved mainly using methods in Data Mining, Machine Learning and Artificial Intelligence. GE Healthcare systems also provide patient monitoring and an aid in making more accurate and certain diagnosis in cardiology using Historical data (“GE Healthcare”, 2016). Other platforms such as webmd.com allows users to create accounts and store data on their current conditions. Webmd.com is then able to study the users uniquely and give them likely outcomes of some symptoms and even drugs they can use to cure the found outcomes. Such systems are however expensive to build with a necessary huge investment in liquid and human capital. Low-income areas can however leverage open-source tools to achieve good public health surveillance while such investments are yet to be possible. This research explores some of the underlying approaches of Machine Learning that are employed in building Biosurveillance systems and concludes on what approach to use and open-source technologies to be employed in proving the feasibility of building public health surveillance systems in low-income countries.

Disease outbreaks usually occur as anomalies. Some anomalies such as measles may be loud enough to trigger the doctor, who in turn cautions the community. However, if for instance only the temperature of patients begins to change to about only a degree Celsius or two (for instance with the case of Ebola), it will be difficult to understand that as a pattern

6

for a possible disease outbreak. With a computer program that constantly studies the patterns of different metric data from patients, it will be easy to report the faintest changes in patterns to incite an action. A system that constantly checks for these anomalies in patients' data will reduce the risk of disease outbreaks by alerting health practitioners at the early stages of the outbreak.

Such systems can either be built to monitor patients separately to warn if their habits or symptoms are leading to a disease, or to monitor a population of patients to find trending patterns that may suggest the health conditions of the population.

This research investigates the feasibility of using open-source technology to forecast diseases more accurately and detect disease outbreaks early in order to save more lives, tailor national expenditure on health towards interventions that matter, and finally reduce national expenditure on disease outbreaks by delivering prevention through early detection instead of cure.

In the tech-driven developed countries, there is an active technological environment equipping the health industry with that helping assistant using approaches in Data Mining and Machine Learning (Artificial Intelligence) The following are some of the works that either tailor their effort towards disease outbreak detection or forecasting diseases. This paper will discuss these researches based on the purpose of research, methods used, and data collection type, and synthesize why some decisions were taken in each paper so as to aid a better experimentation and validation of using a feasible open-source technology.

In detecting anomalies in large multidimensional dataset such as public health data, Kaustav et al used spatial data alongside a unique idea of having a local anomaly detector to identify individual records with anomalous attribute values. Thus, instead of detecting if the entire record is anomalous, separate local anomaly detectors are used for the different

subsets of each record. Two methods were evident; implementation and experimentation. Kaustav et al got the following dataset to experiment with their implementation; PIRS Dataset which has records describing containers imported into country from various ports in Asia, Emergency Department Dataset which consists of real-world dataset containing records of patients visiting emergency departments from hospitals around Allegheny county in 2004, and lastly KDD Cup 1999 Network Intrusion Detection Dataset which consists of Intrusions simulated in a military network environment.

Instead of only spatial or temporal data, George et al used spatio-temporal data. Dataset used are a year worth of patients ED Data provided by the Pennsylvania Department of Health. All patients were anonymously modeled into nodes, with attributes as subsets that described the health conditions of the patient. Each node however had a conditional probability of correlating to one or more diseases, based on the current health conditions of the patient. If more nodes are inferring a particular disease with a total probability greater than the set threshold probability, then there is a high likelihood that the inferred disease is an outbreak. It was however not stated what happens when nodes with the subpopulation tend to have different attributes from the rest of the subpopulation.

Yang et al used social media analysis for diseases surveillance in Beijing. The team of researchers collected social media posts from a China micro-blog. Yang et al implemented two systems based on K-Means (unsupervised) and SVM (supervised), and then experimented with data from Chinese micro-blog to find out which of the two methods produced more accurate results. From their experimentation, SVM produced the most accurate results matching the actual data reported.

Kaundal et al propose that since neural networks and multiple regression are unable to predict value of unknown data points and longer training times, Support Vector Machines

is a better approach to forecasting diseases of plants, putting into account factors such as weather. Historical data on farm diseases over the year 2000 was gathered and correlated to meteorological data. Weather variables such as rainfall, temperature, relative humidity, sunshine and etc. then helped predict the next set of diseases expected based on the weather conditions that are expected, and their correlation and causation to past data. Different approaches including; Artificial Neural Networks, Multiple Regression, and Support Vector Machines were tested on the data. From Kaundal et al, SVM produced the most accurate forecast.

Last but not least, Rana et al disapproved using SVM, Naive Bayes or Random Forest as predictive frameworks for clinical interventions. This is because, the mentioned approaches amalgamate input variables into a single rule for the all their predictions. Clinical interventions however evolve as new diseases and new approaches to solving diseases, or even new understanding of diseases are discovered. Rana et al rather proposed a predictive framework that separates interventions from patient condition. That way, just like in Object Oriented Programming, changes can be made to the separated interventions as they evolve so to understand current trends in patient data in order to make the right predictions—thus, an evolving rule for prediction is created instead of a definite rule.

2.1 Discussion

Insightful information are mined today from analyzing social media posts from large networks such as Facebook and Twitter. Yang et al have proved that it is even possible to get insightful information from not so popular social networks such as Chinese micro-blog weibo.com. Getting rich insights for medical research will however require that a significant percentage of the population under study uses the social network in question. In low-income economies such as Ghana, not everyone uses these networks. Only 20% of Ghanaians have access to the Internet and only 11% of Ghanaians are on the most popular social network, Facebook, as of November 2015 (“African Internet”, 2016). Youth are the main users of these platforms, and in rural and remote areas almost a negligible number of people use these networks. Gathering data from patients or residents will require a different approach than social media analysis.

Kaundal et al in their approach to forecast the next m plant diseases over a period of time, in correlation to weather condition made an interesting point that the time required for an algorithm to learn the training set matters. Kaundal et al chose SVM over other approaches because SVM is able to predict values of unknown data points and also requires less time to learn patterns in training dataset. What if we do not have enough data to serve as training data? Most rural places in Ghana and other low economies have very small population. Getting enough data as a training data may be impossible. Due to the advent of big data, it is however possible to model a Biosurveillance system based on data collected from another region with very similar characteristics such as weather patterns and historical record of disease cases. The system must however be able to cope with changing medical interventions and newly discovered diseases so be always current with actual health trends. SVM approach used by Kaundal et al does not learn continuously, making it a less viable candidate for an area with rapid changes in medical interventions.

This research will therefore prefer an approach that learns continuously to catch up with changes in data patterns.

The idea of computing an anomaly likelihood in any case of a discovered anomaly by Kaustav et al inspires this research in its detection of disease outbreaks. This is to make sure some detected outbreaks are not mere noise in data by computing a probability normal chart of historical anomalies to compare with detected anomaly. A true anomaly has anomaly likelihood above chosen threshold. Threshold in this case depends on the sensitivity of the data and the spatial representation of the data.

After analyzing the researches above, HTM (Hierarchical Temporal Memory) Algorithms implemented by Numenta seemed feasible due to the following characteristics:

- Anomaly Detection
- No need to store data
- Online prediction
- Continuous learning
- In-built clustering mechanism

(“Applications of Hierarchical”, 2015)

2.1.1 Limitations of these Theses

These theses may not work appropriately in low-income countries due to a number of reasons including: technology, availability of data, and cost of implementation and maintenance.

2.1.1.1 Technology

In low-income countries, technology is not very accessible (“Global forum”, 2010), therefore building a system that requires different separate components may be difficult. All the approaches discussed above solve only one problem, thus either forecast diseases over the next period of time, or detect disease outbreaks with the exception of Yang et al whose research depend on social networks. This means implementing a system that requires separate implementations for each analysis will need more technology. Thus, a separate system for disease outbreaks, and a separate system for forecasting, and so on and so forth. It could in the long run, as data increases, require these systems run on different clusters of storage with redundant data stored across separate clusters. For instance in the case of George et al, all nodes in the network (each node representing a patient) retain values in sub networks that maintain a description or state of the patient. Values are referred to in making inferences. Therefore if another system is to be built to manage for instance forecasting, some similar data may have to be stored separately for the new system.

HTM Algorithms do both prediction and anomaly detection with a single model (clustering is also achieved) (“Science of Anomaly”, 2015). This means instead of two models, a single model is required to achieve both forecasting and anomaly detection. The different functionalities within a single model cooperate in learning patterns. Prediction for instance helps with anomaly detection by giving a clue to the anomaly detector that a data point or pattern of data was unpredictable hence requires a further check by the anomaly detector. This makes HTM systems viable for an implementation that requires more than a single functionality. Also as new patterns are learned, all different learning mechanisms are aware of the new patterns and hence collaborate in analyzing the next set of data (“Application of Hierarchical”, 2015). For instance, after the anomaly detector finds an anomaly, the entire system learns the new pattern on subsequent appearances. If the entire

system recognizes the pattern or value is no more anomalous, the predictor begins to predict it. If the earlier anomalous value is no more found in future data, HTM “forgets” it as its conditional probability of occurring reduces incrementally. This makes HTM recognize it on its next appearance as an anomaly (“Science of Anomaly”, 2015).

2.1.1.2 Availability of Data

The approaches used by all the method discussed in related papers are unsupervised learning methods. This means they require training data to build models and rules to help with either prediction or anomaly detection. From evidence in collecting medical data in Ghana during this research, it is unlikely to use a supervised method that requires large amount of data to learn patterns and generate rules for analyzing future data. This is not a problem however since with the advent of big data, data from countries with similar characteristics such as historical health conditions can be used to model a supervised system. The problem however is that as health issues and their interventions evolve, baseline distribution of historical data become misleading. From the table below, it is clear that only little amount of data was found. For the communities that were studied (community4 and community17), data was multiplied. However, since MATE is unsupervised and does continuous learning, it is seen in experiments below that MATE learned new patterns of data that were parsed and adjusted its predictions in response to the current baseline distribution. MATE therefore proved to be self-correcting as it learns continuously without any human intervention.

Table 2-1 Communities and their data size

Community	Number of records in data
community1	377
community2	561
community3	391
community4	757
community5	702
community6	611
community7	89
community8	398
community9	253
community10	268
community11	519
community12	338
community13	419
community14	343
community15	171
community16	242
community17	782
community18	150

If data is not “messy” or noisy, with well define sequentially repeating patterns such as a sign function, such amount of data could work. Health data unfortunately is evolves each day with new patterns. It will therefore be necessary to have huge data for a system to repeatedly learn as its training set. Otherwise, a very good unsupervised learning approach is better since it learns new patterns incrementally as data evolves with evolving medical

14

interventions and disease outcomes instead of manually computing new baseline distributions.

2.1.2 Cost of Implementation and maintenance

The approach of Rana et al to separate interventions from patients' records makes it easier to add new rules to their system by updating existing interventions. For all other researches mentioned, as new interventions are found, there is the need to repeat earlier steps in implementation and testing processes to be sure new trends in the expected functionality of the system are achieved. Maintenance can be very expensive as according to ("Cost to develop", 2016) it was reported in building a single state surveillance system for clinical laboratories, it costed New York \$10.55 million for development over 5 years and \$485,000 for maintenance between 2006 and 2013. With HTM algorithms however, new patterns are automatically learned as time goes. This means there is little or no cost in maintaining MATE aside scaling after it is completely built. MATE also does not need to store input data, hence the need for storage is only the choice of the implementer, but not a requirement.

2.1.2.1 Cost of running a model for each community in Ghana simultaneously

Since there are 275 constituencies in Ghana (Kuruk, 2012), assume each constituency has 60 neighborhoods (Accra has 59). Number of HTM models required to run is therefore 16,500, from 60×275 . 450 instances of MATE can run efficiently on an i7 computer with 8GB RAM ("HTM Engine", 2015). This is to calculate the worst case scenario, otherwise more models of HTM can run simultaneously. 450 models is chosen to reduce impact on hardware. If each computer runs 450 models, a model for each neighborhood, then 36 computers are required (from $16,500 / 450$). Cost of each computer (could be refurbished)

is estimated at GhC 2,000. Total cost of running a model for each community in Ghana simultaneously is therefore about GhC 72,000 for hardware. There is GhC 0.00 cost for software since Ubuntu 15, the OS on which MATE runs, is open-source. All other libraries used are open-source. Aside hardware, the only cost involved is payment for developers to implement and maintain MATE.

Chapter 3 Method

From all the papers analyzed during this research, implementation and experimentation were a common pattern. Cooper et al went further to simulate their system. This research will therefore be primarily be based on implementation and experimentation with available dataset.

3.1 Implementation

This research chose implementation as an approach because it has been the recurrent theme in almost all research papers explored in the literature review. This is as a result of the how delicate health matters are. It is appropriate regardless what approach any health research uses to implement concept and test thoroughly before use, so an intended solution does not turn out to be a hazard.

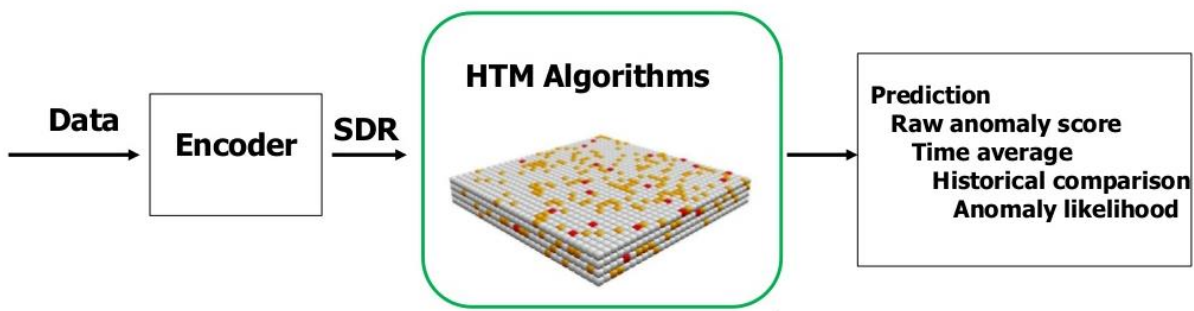


Figure 3-1 Architecture of HTM

Figure 3-1 above shows the structure Architecture of HTM used in this research. HTM is inspired by the structure and functioning of the mammalian neocortex (“Science if Anomaly”, 2015). Terms such as *cells* will be used to represent abstract structure in HTM that are supposed to abstract the brain cell or a neuron. For a data point in the brain, some brains cells are active while others are inactive, the distribution of active cells versus inactive cells gives each information a unique representation. This representation is

achieved in HTM by using Sparse Distributed representations, where 1s represent active cells and 0s represent inactive cells.

For HTM to work, all input data are first converted into Sparse Distributed Representations (binary vectors) by 4 different types of encoders at the moment. Encoders include: *scalar*, *categorical*, *datetime*, and *GPS* encoders. HTM Algorithms are then fed with the SDR, making the algorithms understand the semantic meaning of the inputs (since all inputs are binary vectors). HTM Algorithms include the CLA (Cortical Learning Algorithms) Classifiers, Anomaly and Anomaly Likelihood Algorithms, Anomaly classifiers, and KNN Classifiers. These algorithms work together to form a single “brain” of interacting algorithms.

To model the data in this research, categorical encoder was used to model the list of diseases streaming into MATE. *datetime* encoder was used to model the timestamp for each data point, and GPS encoder was used to model the location of the hospitals. Below is the architecture of MATE.

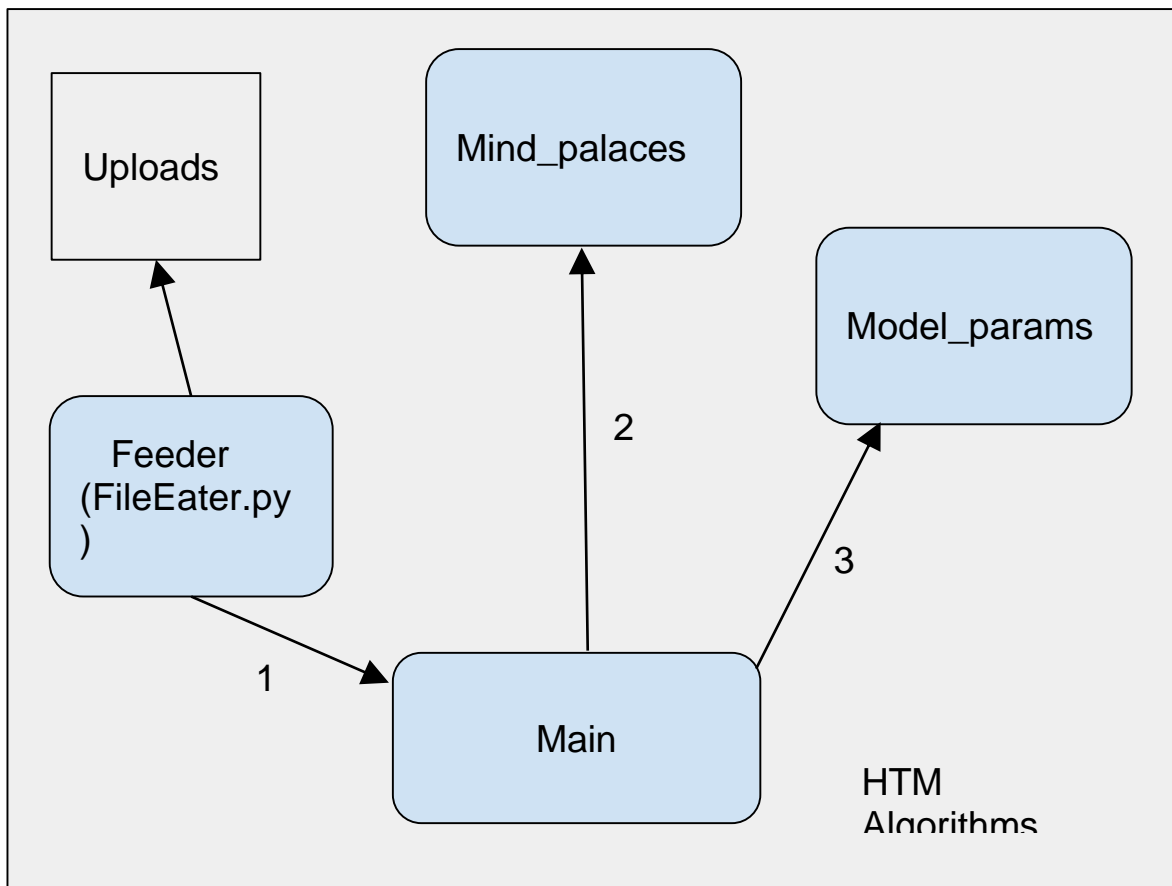


Figure 3-2 Architecture of MATE

Figure 3-2

MATE has four main layers. The *feeder* layer, the *mind_palaces* layer, *model_params* layer and the *Main* layer. Each layer represent a directory that completes specific task in the cycle of MATE. A scenario is used below to explain how the various layers work together.

3.1.1 How MATE works

Assuming a nurse uploads a csv file of disease reported over a week to an arbitrary folder called *uploads*. The Feeder layer is caused to react to the upload. The feeder reads the file to know what community is it coming from. Based on the community and the sort of data it is, the Feeder layer runs a specific instance of main program in the *Main* folder. The running instance first checks the *mind_palaces* folder if there is already an existing model of the

submitted data. If there is, the program simply reloads the model and reads the new data to make new predictions, find anomalies, and record new patterns. It then updates the existing model of the data in the *mind_palaces* folder.

If the model for that data does not exist, the running main program goes to the *model_params* folder to find a specific configuration for the kind of data submitted. It then uses the configuration to create a model of the data and runs through each data point to learn new patterns. The model created for the first time is then stored in the *mind_palaces* folder. Therefore the next time same kind of data is submitted from the same community, the running main program simply needs to reload the stored model to continue learning.

3.1.2 How Prediction works

Prediction in HTM uses KNN means over sparse distributed data to predict next expected data points. The diagrams below are used to explain how predictions are achieved from HTM algorithms. Figure 3-3 shows a single dimension of a layer in HTM when first data point is parsed.

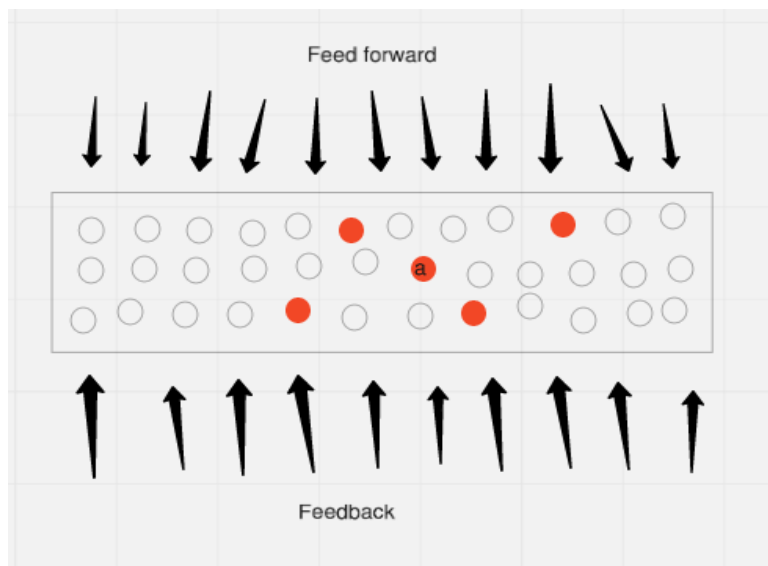


Figure 3-3 A layer of cells in HTM for first data point

For each data point parsed into MATE, Hierarchy Temporary Algorithms in MATE acts as actual hierarchies with different layers as found in the mammalian neocortex. The hierarchies are populated by units of abstraction of cells in the neocortex. Before a data points gets into the hierarchies, it is first parsed through two external layers called spatial and temporal poolers. Spatial and temporal poolers first convert the data point into a Sparse Distributed Representation, which is literally a binary vector of mostly 0s and few 1s to represent the data point. In the mammalian neocortex, this is replaced by brain cells that are either active or inactive. Active cells are represented by 1 and inactive ones as 0. Temporal and Spatial poolers also stamp the data point with a time and spatial representation. The spatial and temporal pooler serve mainly as the feed-forward part of the diagram above. They prepare each data point before it gets to the learning layer of HTM hierarchies. At the learning layer, cells depicted as tiny circles are either active or inactive. Active cells are colored as orange and represent the data parsed. The diagram above shows the case of initial data point when MATE has not seen any earlier data, therefore the feedback region is clueless and could not predict any values. The diagram below shows the nature of cells when MATE is already familiar with the data.

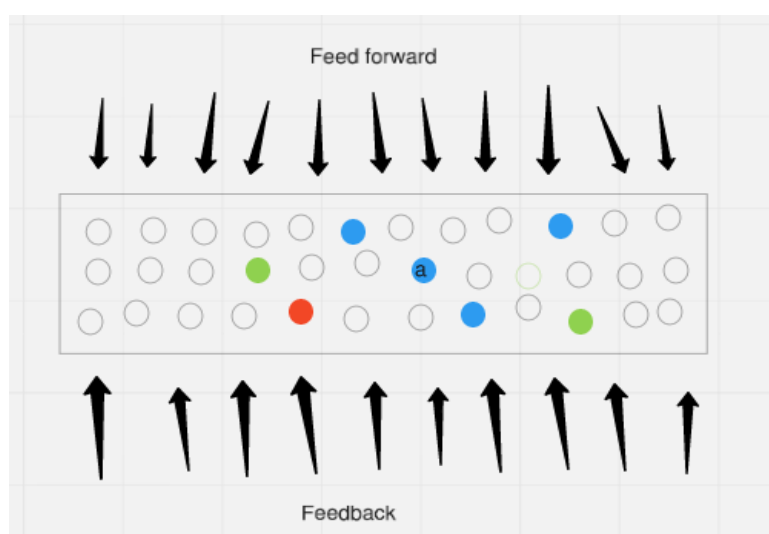


Figure 3-4 A layer of cells in HTM when analyzing familiar data points

As MATE gets introduced to data, each value in data is converted to a binary vector. This means “a” could be represented as “0000000000001000000000000000000001”. The 0s and 1s are used to make some cells active and others inactive. When a cell is active, it uses the K-Means algorithm to find other active cells that are closest to it, and make a connection to them. To remind itself about the connections it just made, the active cell keeps a map of all its connections to other cells and how strong their connection was using a proximity-factor. In effect, all the active cells have together contributed to creating a particular state. The learned pattern is pushed to lower layers in the hierarchy, where similar representations are classified by the CLA (Cortical Learning Algorithms) classifiers. The next time patterns similar to the learned ones are introduced into MATE, the feedback layer is able to predict which cells will be active and which ones will not be active. If the cells predicted to be active are truly activated by the input data, then the prediction is perfect. Otherwise if there are more active which were not predicted, MATE recognizes that data point as an anomaly. Thus, an anomalous data point is that which has a greater number of active cells that were not predicted. The blue circles above show active cells that were predicted, the orange cell is an active cell that was not predicted, and the green is a predicted cell that is not active. Since MATE understands the semantics of the input data, it is able to recognize the above as not anomalous since most of its active cells were predicted.

The representation above is only two-dimensional for easy understanding. True nature of HTM is three-dimensional space-time. Time-stamped data points give data its temporal feature. In this research, instead of correlating dataset with external influences such as weather, time is acknowledged to have strong correlation to many external influences including weather.

3.1.3 How Anomalies are detected by MATE

Both Spatial and Temporal patterns are learned by HTM Algorithms, hence discovering unexpected patterns is augmented by both spatial and temporal inferences by the spatial pooler and the temporal pooler respectively. Since temporal nature of anomalies are considered, a known pattern can still be an anomaly if it occurs at an unexpected time.

Two algorithms in HTM detect anomalies: the Anomaly algorithm and the Anomaly Likelihood algorithm.

3.1.3 Raw Anomaly Score

In HTM, anomalies are computed by finding the fraction of active columns or cells that were not predicted. If raw anomaly score is 0, then both spatial and temporal patterns were perfectly predicted. If raw anomaly score is 1, then spatial and temporal patterns of the new data were totally unrecognized. Below is a mathematical formula for the computation of raw anomaly score.

$$\text{rawAnomalyScore} = |A_t - (P_t \cap A_t)| \div |A_t|$$

A_t = Actual cells

P_t = Predicted cells

3.2 Dataset

Grameen Foundation, Ashesi-mHealth and Numenta provided data for this research.

3.2.1 Data from Numenta

Data from Numenta's sample application *hotgym* was used to test if HTM system installed works as expected. The *hotgym* data is real energy consumption data from a gym in Australia, which simply contains a timestamp and float value for energy consumption. All

expected anomalies pointed out by Numenta were discovered when experimented with MATE. Also new patterns were learned as MATE run through the *hotgym* data as expected.

3.2.2 Data from Ashesi-mHealth Project

Ashesi-mHealth provided record of vaccines used in treatment of patients. Data is evenly distributed. With forecasted vaccines, we can infer the sought of diseases that could occur. Also with changes in patterns of the usage of vaccines, we can detect signs of outbreaks.

3.2.3 Inpatients reported cases from Grameen Foundation

Anonymized dataset of inpatients and their reported cases and location of their hospital was provided by Grameen Foundation Ghana. The dataset was cleaned by first sorting all records by location. The different communities were then exported to different files. Finally all false reported cases were removed from each community. The final dataset was a list of files with true reported cases by anonymized patients from same communities.

3.3 Experiments

Three different experiments were conducted with mate to ascertain the following:

- How accurate are the predictions made by MATE if:
 - Data is highly biased (thus, huge percentage of data are a particular value)
 - Data is evenly distributed
- How quick does MATE recognize anomalies (Disease outbreaks)
- How quick does MATE recognize new patterns in data to realize an expected anomaly is now a pattern?

3.3.1 Experiment 1(a)

Measuring accuracy of MATE using skewed data.

Dataset: List of reported diseases by patients over the period of 2012-2012

Source of data: Grameen Foundation Ghana

Community: Community 4

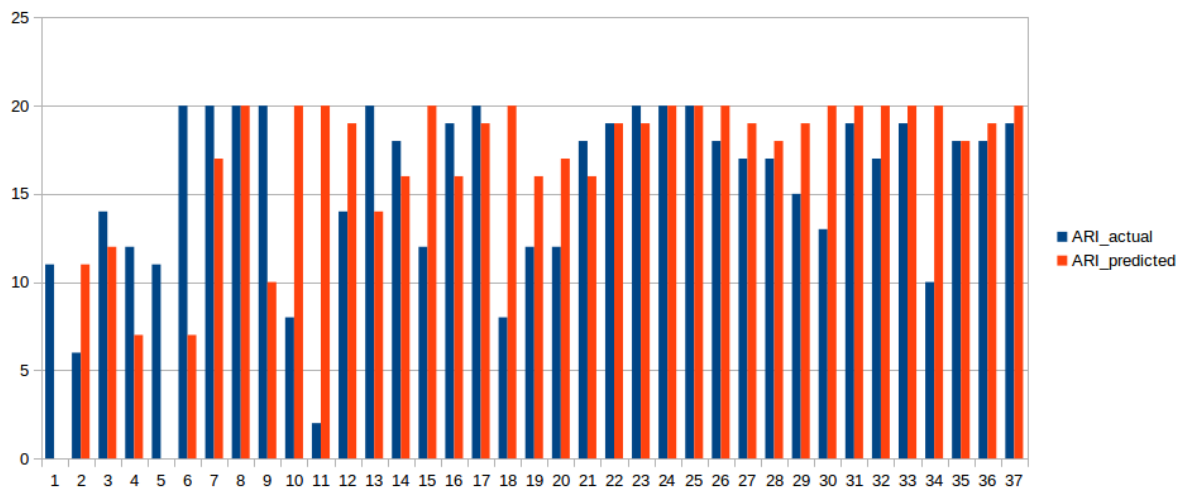


Figure 3-5 ARI actual vs. ARI predicted in skewed data, community4

Predicted versus actual cases of ARI. Blue bars represent actual reported cases of ARI in 20 every 20 reported cases of diseases. Red bars represent predicted cases of ARI in the next 20 reported cases of diseases.

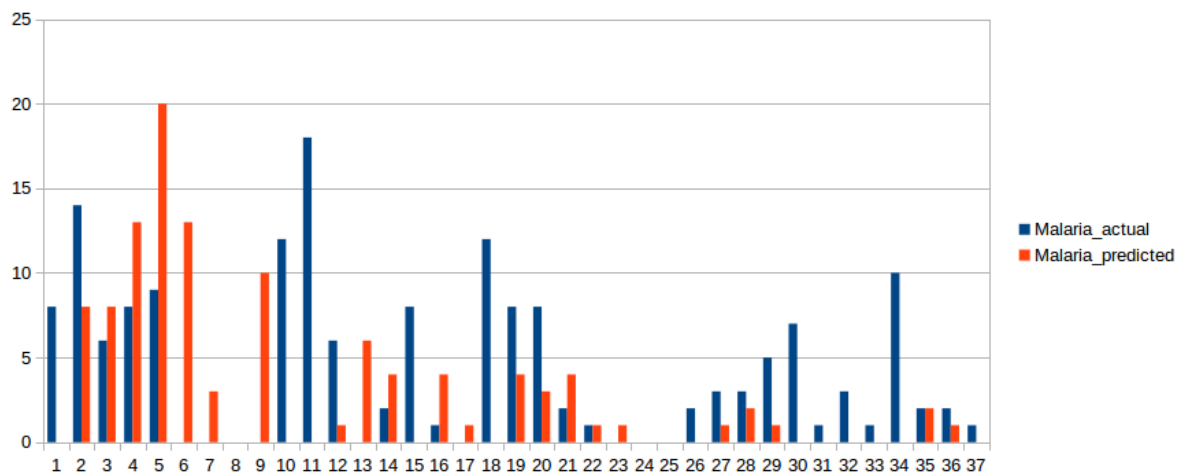


Figure 3-6 Malaria actual vs. Malaria predicted in skewed data, community4

Predicted versus actual cases of Malaria. Blue bars represent actual reported cases of Malaria in 20 every 20 reported cases of diseases. Red bars represent predicted cases of Malaria in the next 20 reported cases of diseases.

Community: Community 17

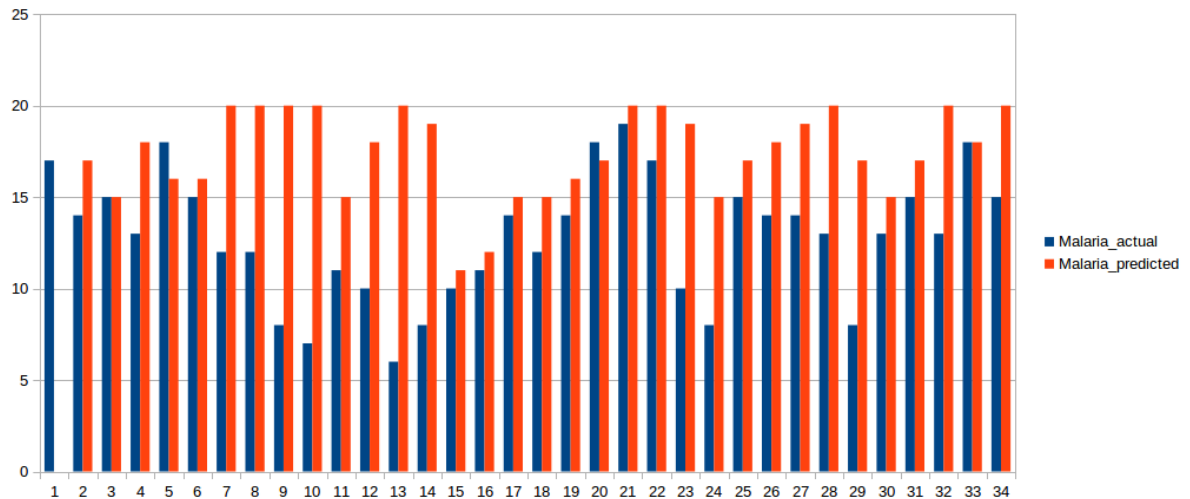


Figure 3-7 Malaria actual vs. Malaria predicted in skewed data, community17

Predicted versus actual cases of Malaria. Blue bars represent actual reported cases of Malaria in 20 every 20 reported cases of diseases. Red bars represent predicted cases of Malaria in the next 20 reported cases of diseases.

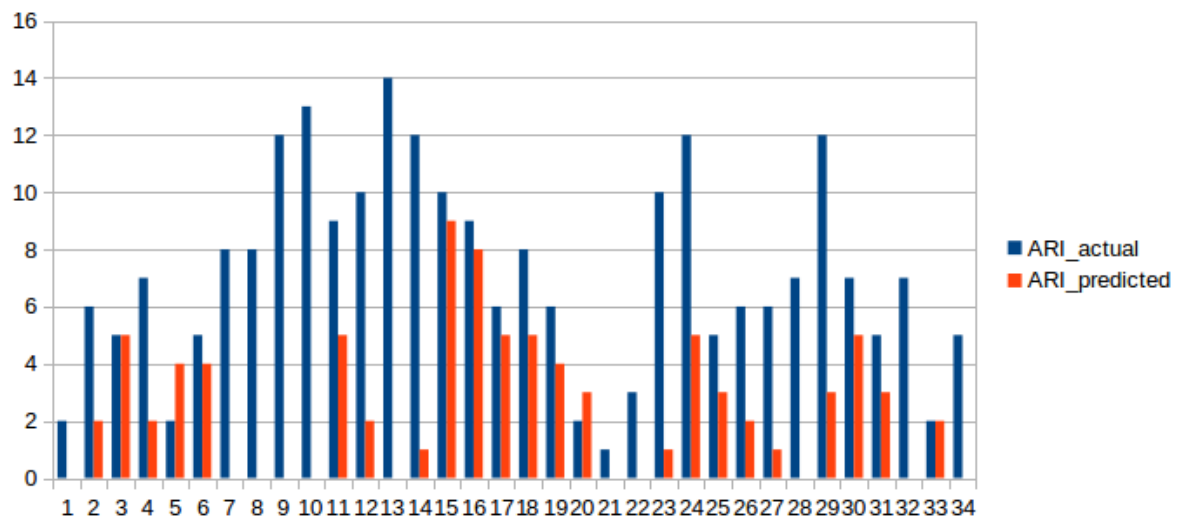


Figure 3-8 ARI actual vs. ARI predicted in skewed data, community17

Predicted versus actual cases of ARI. Blue bars represent actual reported cases of ARI in 20 every 20 reported cases of diseases. Red bars represent predicted cases of ARI in the next 20 reported cases of diseases.

3.3.1.1 Observations

For both communities, one disease is at least about two-thirds of all cases. In community⁴, Malaria occurred on 163 times while ARI reports added up to 589 (ARI cases form 78.34% of all cases). In community¹⁷, Malaria was reported 443 times while ARI was reported only 246 times (Malaria cases form 64.23% of all cases).

3.3.2 Experiment 1(b)

Measuring accuracy of MATE with an evenly distributed data in forecasting the next **20** vaccines to be used in a hospital.

Dataset: List of vaccines required by patients over the period of 2012-2015. 20 different vaccines were used at different times. The experiment seeks to find out if MATE can forecast 20 different items though the various columns depend on each other by conditional probabilities.

Source of Data: Ashesi-mHealth program.

After running MATE over data containing to predict the next 20 vaccines to be used in a hospital, the following were discovered after running through $n * 20$ rows of data. Thus, MATE reads a series of 20 data points and learns the spatial and temporal patterns with the 20 records. Below are charts comparing actual distributed of data, in this case the number of times the vaccines BCG and Vitamin A (6month) were actually needed versus the predictions by MATE.

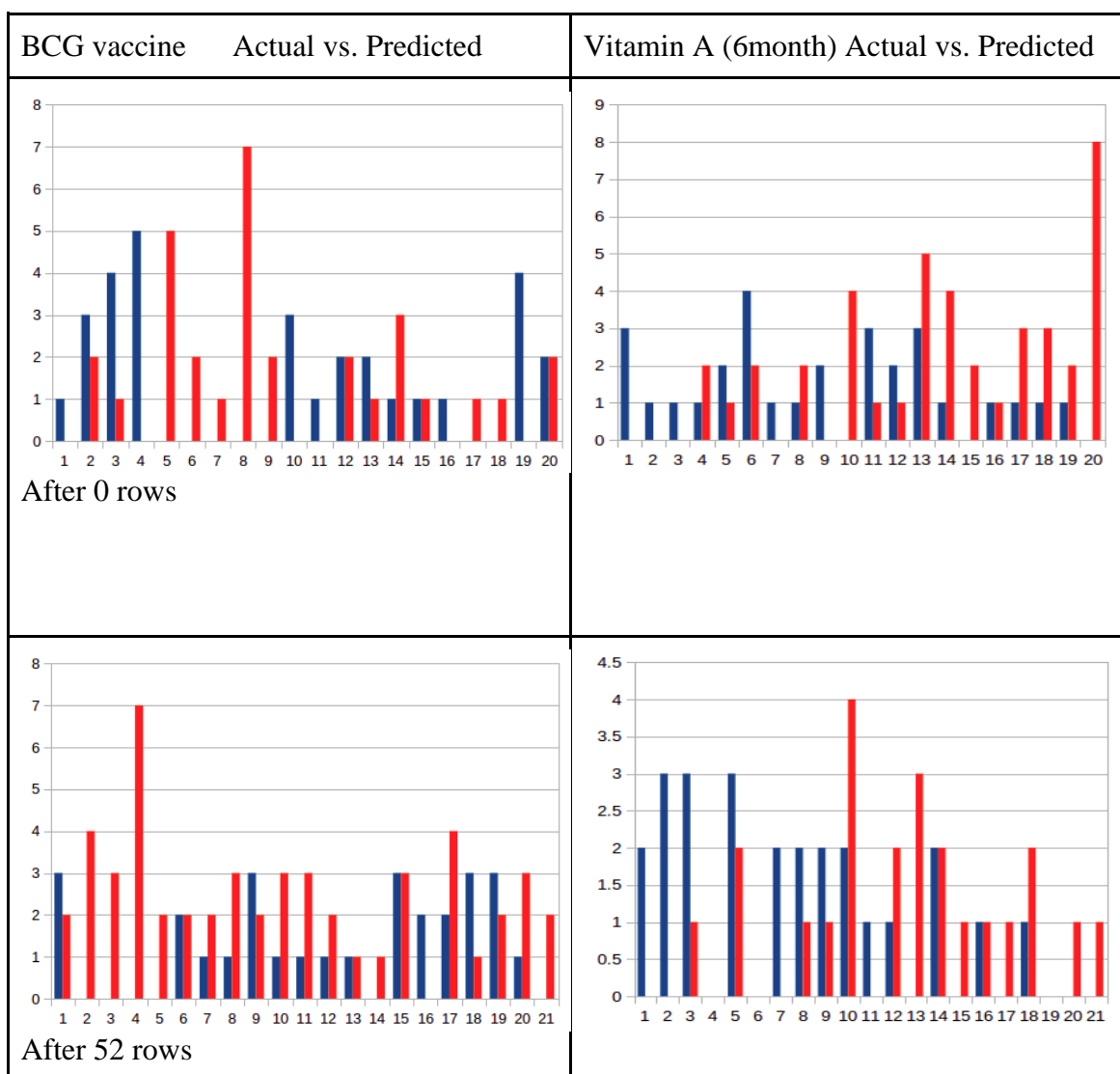
Key

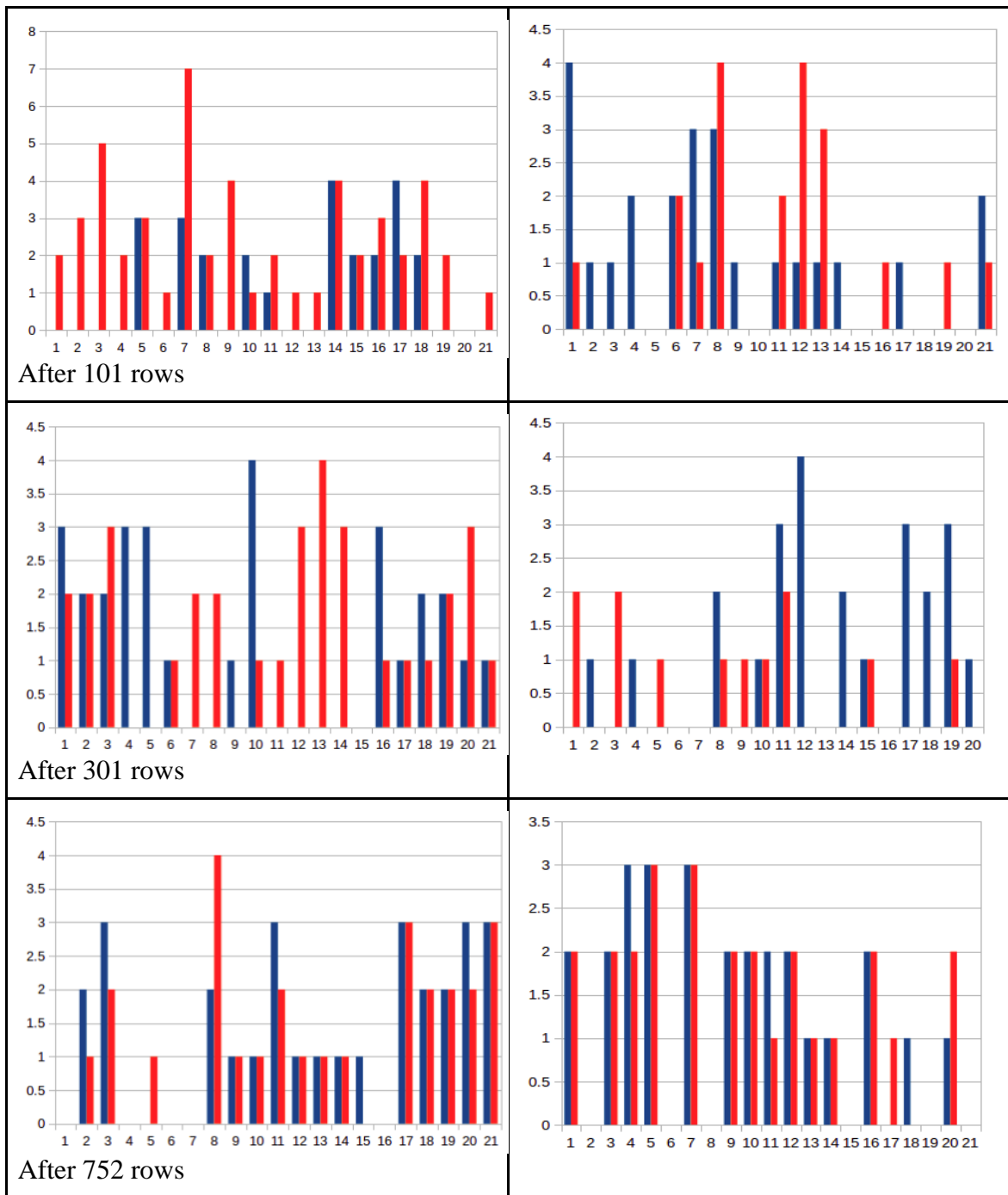
Blue bars represent actual number of times vaccine was used.

Red bars represent the predicted number of times vaccine would be required.

Each row expressed in diagrams below is equivalent to 20 data points since MATE is learning patterns in 20 data points at a time.

Table 3-1 Good predictions with increasing accuracy in evenly distributed data







3.3.2.1 Observation

Accuracy was only counted if and only if actual number of occurrence of a vaccine is equal to the predictions in the forecast. Thus, only a 100% accurate forecast was counted.

- Just as expected MATE could not make any predictions when reading its first 20 rows of Data (Seen in first graph on top left). This is because since MATE is

unsupervised in its learning approach, it knows absolutely nothing about the data. After the first 20 rows however, MATE was able to find new patterns and recognize some values to expect in the future.

- After 752 rows of data, each row equivalent to 20 data points, it was realized that MATE made very good predictions with high certainty. Below is table showing accuracy achieved after **x** number of rows were learned by MATE.

Table 3-2 Accuracy of MATE data increases

After x rows	Accuracy for BCG	Accuracy for Vitamin A(6month)
X = 752	66.66%	80%
X = 903	71%	80%
X = 1004	80%	90%
X = 1174	85%	76%

- It was observed that the anomaly introduced in input data to support the next experiment affected the accuracy on predicting Vitamin A (6month) but not BCG. From the predictions of Vitamin A (6month) before 1174th row, it is evident that Vitamin A (6month) has a higher conditional probability of occurrence than BCG (Vitamin A (6month) occurs 1121 times while BCG occurs 816 times in data). A little change in the pattern of data therefore is expected to have affected affect Vitamin A (6month) more since Vitamin A (6month).

3.3.3 Experiment 2(a)

How quick does MATE recognize anomalies (Disease Outbreaks)?

Dataset: List of vaccines required by patients over the period of 2012-2015.

Source of Data: Ashesi-mHealth program.

New vaccine was added to data at row **1004** when MATE had learned most of the patterns in the data. This experiment was meant to find how responsive MATE is when new data are included into data MATE is already familiar with.

The following charts explain how MATE first recognizes all data points as anomalous until it has seen them again.

Figure 3-9 shows first set of data being anomalous, and as MATE learns it begins to recognize the patterns and vaccines parsed. [First 500 rows]

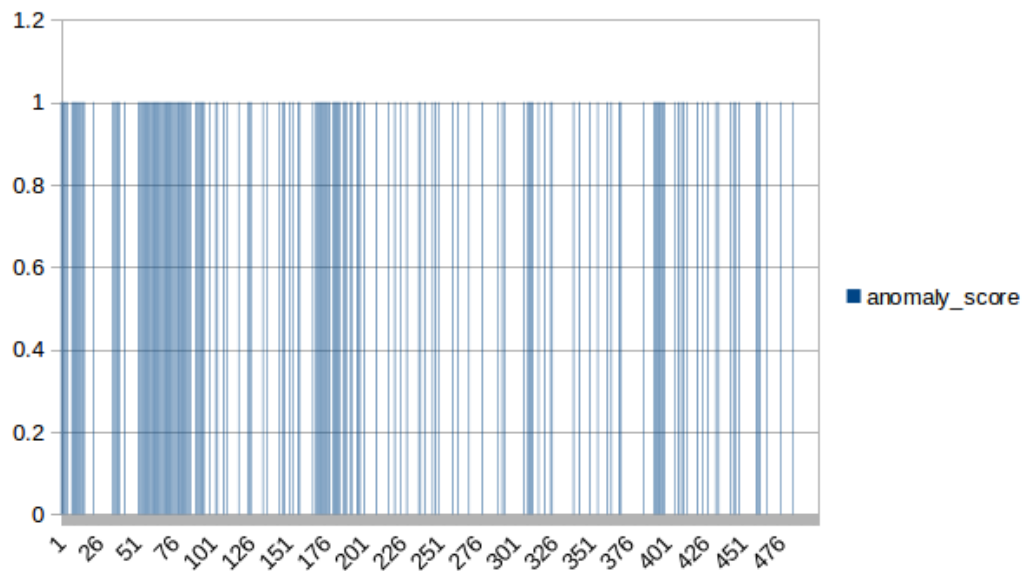


Figure 3-9 Many anomalies found in new dataset

Figure 3-9 shows fewer anomalies after 1000 rows of data. At this point MATE was familiar with all vaccines but some patterns (spatial and temporal patterns) were still anomalous.

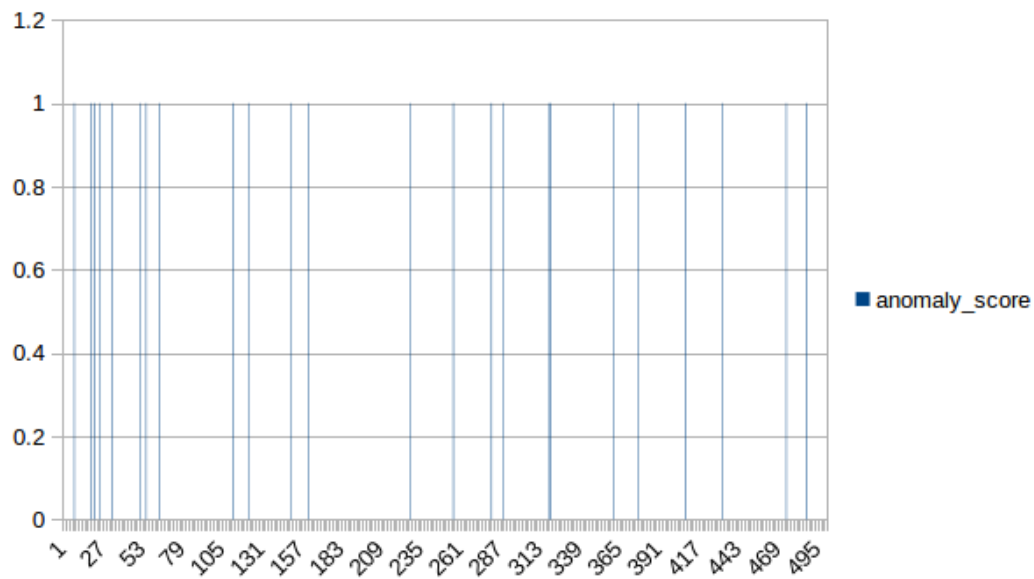


Figure 3-10 Fewer anomalies found as MATE learns

Figure 3.3.1.3 shows not a single anomaly after reading 3336 rows of data

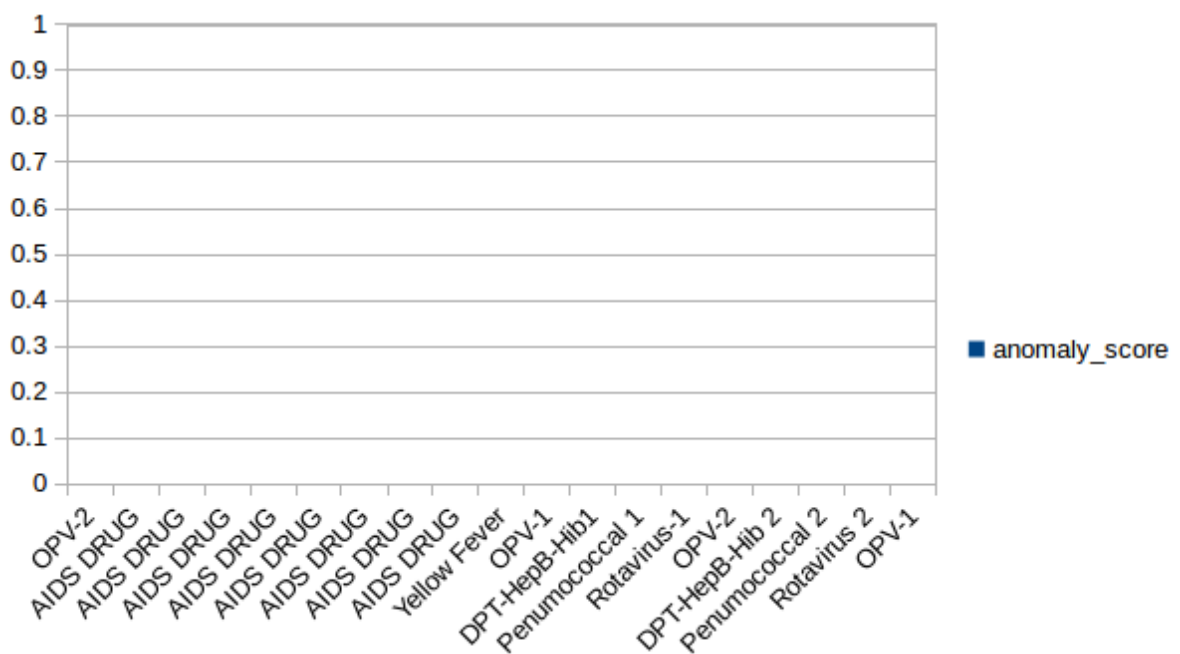


Figure 3-11 No anomalies are found after MATE learns enough patterns

Figure 3-11 shows MATE recognizes an anomaly when new vaccine (AIDS DRUG) was parsed.

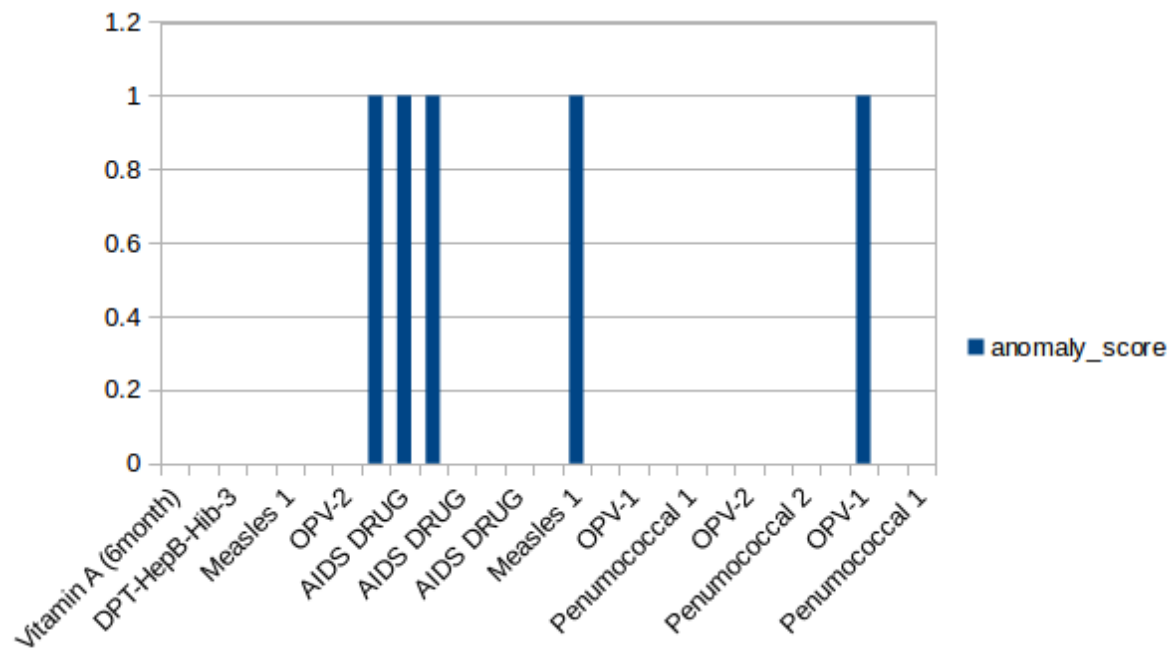


Figure 3-12 Anomalies found on their first occurrence

Observation

It was observed that when MATE was first introduced to new dataset, all data points were reported as anomalous data points. As MATE got familiar with the data, the reports reduced. If compared to 1(b), it is clear that as MATE learns, prediction accuracy increases as MATE recognizes fewer anomalies. Figure 3-11, almost all patterns in data have been learned by the different HTM models running in MATE. At this points, we can trust all reports from MATE.

When almost all patterns were already learned, anomaly was introduced to investigate how quickly MATE will detect it. A vaccine named AIDS DRUG was introduced. It can be seen in Figure 3-12 that MATE detected the anomaly with high certainty.

3.3.4 Experiment 2(b)

How much fast does MATE learn new patterns in new data?

Is was seen from above that MATE recognizes all new values as anomalies, and then as the same values appear, their patterns are learned. When MATE corrects itself that a value is not anomalous, the new value gets to be predicted once MATE learns its patterns of occurrence. Below is a follow up of **Experiment 2_subA**. It will be noted that MATE recognized later as it got more “AIDS DRUG” vaccines, and hence stopped reporting “AIDS DRUG” as an anomaly

Fig. shows MATE still quite uncertain if “AIDS DRUG” is anomalous.

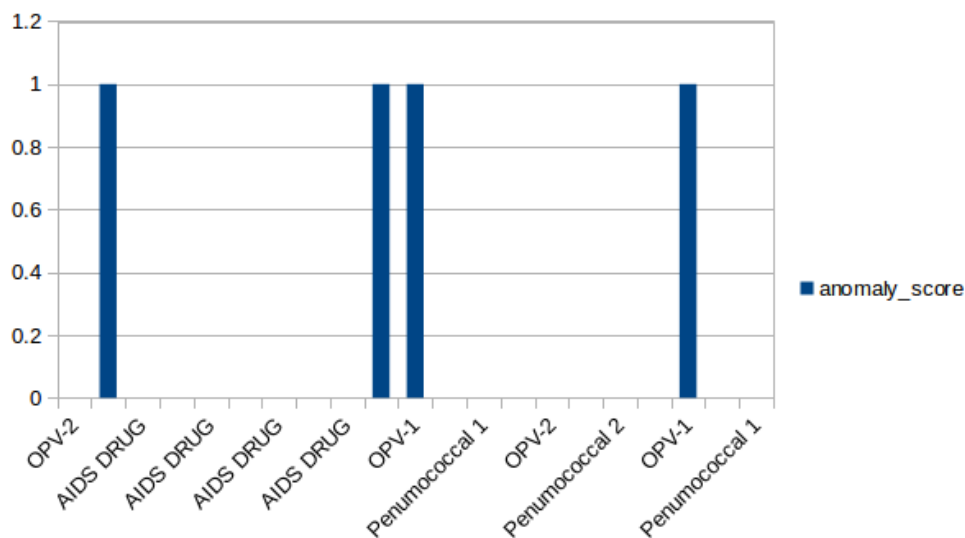


Figure 3-13 Anomalies discovered again not as values but as patterns

Figure 3-13 shows after seeing “AIDS DRUG” appear three times, MATE did not flag any anomaly report.

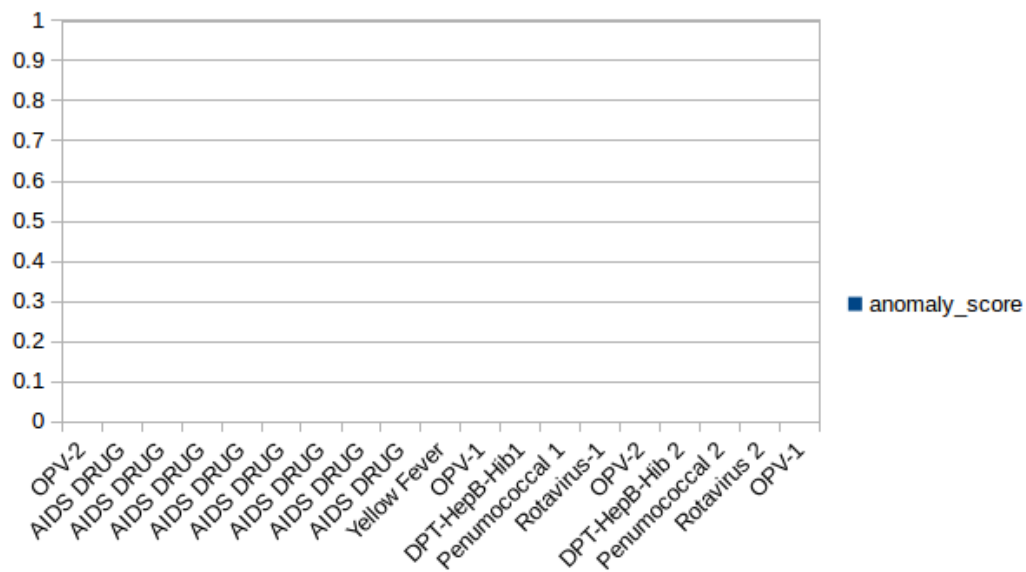


Figure 3-14 Mate does not found earlier anomalous data as anomalous after learning patterns of the anomaly.

Figure 3-14 shows MATE not actually predicting “AIDS DRUG” after seeing it for the third sequence and MATE did not find it as an anomaly.

3827	timestamp	actual	predicted	anomaly_score
3828	2014-02-10 16:40:00	BCG	Penumococcal 2	0
3829	2014-02-11 08:00:00	DPT-HepB-Hib 2	Penumococcal 2	0
3830	2014-02-11 08:20:00	Penumococcal 2	Penumococcal 2	0
3831	2014-02-11 08:40:00	Rotavirus 2	Yellow Fever	0
3832	2014-02-11 09:00:00	OPV-2	AIDS DRUG	0
3833	2014-02-11 09:20:00	OPV-2	OPV-2	0
3834	2014-02-11 09:40:00	DPT-HepB-Hib 2	AIDS DRUG	0
3835	2014-02-11 10:00:00	Penumococcal 2	AIDS DRUG	0
3836	2014-02-11 10:20:00	Rotavirus 2	Penumococcal 1	0
3837	2014-02-11 10:40:00	Vitamin A (6month)	Measles 1	0
3838	2014-02-11 11:00:00	OPV-2	DPT-HepB-Hib1	0
3839	2014-02-11 11:20:00	DPT-HepB-Hib 2	Vitamin A(18 month)	0
3840	2014-02-11 11:40:00	Penumococcal 2	AIDS DRUG	0
3841	2014-02-11 12:00:00	Rotavirus 2	OPV-2	0
3842	2014-02-11 12:20:00	BCG	AIDS DRUG	0
3843	2014-02-11 12:40:00	OPV-0	AIDS DRUG	0
3844	2014-02-11 13:00:00	OPV-2	OPV-3	0
3845	2014-02-11 13:20:00	DPT-HepB-Hib 2	OPV-3	0
3846	2014-02-11 13:40:00	Penumococcal 2	Vitamin A (6month)	0
3847	2014-02-11 14:00:00	Rotavirus 2	OPV-3	0
3848	2014-02-11 14:20:00	OPV-2	OPV-2	0
3849	2014-02-11 14:40:00	DPT-HepB-Hib 2	Measles 1	0
3850	2014-02-11 15:00:00	Penumococcal 2	DPT-HepB-Hib1	0

Figure 3-15 MATE begins to predict earlier detected anomaly after learning enough patterns in the occurrence of the anomaly

3.3.4.1 Observation

After not seeing “AIDS DRUG” appear since row 11147, MATE finally stopped predicting occurrence of AIDS DRUG after its last prediction on row 13589. This result shows how MATE can automatically adjust to changes in medical interventions in order to model itself based on current patterns in data. As discussed in Chapter 2.1, data from a region with similar characteristics can be used in modeling a learning system, however if the system fails to adjust to changes in characteristics of the actual region it is used, then its results will be misleading. MATE proves to be viable in such conditions where foreign data is needed to model another place. MATE’s ability to recognize patterns, and “forget” irrelevant patterns makes it viable in such regions where foreign data is necessary.

Chapter 4 : Conclusion

From the results of this research, it is evident that low-income countries can use Machine Learning techniques to improve public health surveillance, to tackle such problems as forecasting of diseases and, disease outbreak detection. Although supervised learning models have been preferred in considerable number of researches for public health surveillance, this research has proven that with HTM algorithms, we can equally achieve the accuracy of supervised learning techniques, and even do better as new patterns in medical data are continuously learned by HTM algorithms without any human intervention to update the models with new baseline distribution in data.

Cost of implementing MATE, a system that leverages HTM Algorithms for public health surveillance in Ghana is only at about ₵ 72,000 for hardware. All libraries software used for this research are open-source. Although ubiquitous computing techniques require huge infrastructure sometimes, with the right algorithms and the advent of open-source technology, many problems can be solved cheaply and creatively. One of the problems, as this research has lucidly explained, is public health surveillance in low-income countries.

This research will continue to investigate deeply HTM Algorithms and how to leverage other used algorithms to improve MATE. The researcher will promote this work by presenting his findings at different health institutions in Ghana to promote the use of ubiquitous computing methods such as Machine Learning to improve healthcare.

References

- Gregory F. Cooper, Denver H. Dash, John D. Levander, Weng-Keen Wong, William R. Hogan, and Michael M. Wagner. 2004. Bayesian biosurveillance of disease outbreaks. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (UAI'04). AUAI Press, Arlington, Virginia, United States, 94-103.
- Kaustav Das, Jeff Schneider, and Daniel B. Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '08). ACM, New York, NY, USA, 169-176. DOI=<http://dx.doi.org/10.1145/1401890.1401915>
- Rana Santu, Gupta Sunil, Phung Dinh, Venkatesh Svetha. (2015). A predictive framework for modeling healthcare data with evolving clinical interventions. *Statistical Analysis and Data Mining: The ASA Data Science Journal Statistical Analy Data Mining*, 8(3), 162-182. doi:10.1002/sam.11262
- Yang, N., Cui, X., Hu, C., Zhu, W., & Yang, C. (2014). Chinese Social Media Analysis for Disease Surveillance. 2014 International Conference on Identification, Information and Knowledge in the Internet of Things. doi:10.1109/iiki.2014.11
- Kaundal, R., Kapoor, A. S., & Raghava, G. P. (2006). Machine learning techniques in Disease forecasting: a case study on rice blast prediction. *BMC bioinformatics*, 7(1), 1.
- Wong, W. K., Moore, A., Cooper, G., & Wagner, M. (2003, August). Bayesian network anomaly pattern detection for disease outbreaks. In *ICML* (pp. 808-815).
- Surpur, C. (2015, February 23). Applications of Hierarchical Temporal Memory (HTM). Retrieved April 17, 2016, from

<http://www.slideshare.net/numenta/applications-of-htm-workshop>

Purdy, S. (2015, February 23). Science of Anomaly Detection. Retrieved April 17, 2016, from <http://www.slideshare.net/numenta/science-of-anomaly-detection>

Costs to develop and maintain a state biosurveillance system: The New York example. (n.d.). Retrieved April 17, 2016, from <http://www.cidrap.umn.edu/practice/costs-develop-and-maintain-state-biosurveillance-system-new-york-example>

Hawkins, J. (2014, October 17). Learning Content. Retrieved April 17, 2016, from <http://numenta.com/learn/principles-of-hierarchical-temporal-memory.html>

IBM Watson Health: Welcome to the New Era of Cognitive Healthcare. (n.d.). Retrieved May 01, 2016, from <http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/>

GE Healthcare. (n.d.). Retrieved May 01, 2016, from http://www3.gehealthcare.com/en/products/categories/diagnostic_ecg

Africa Internet Stats Users Facebook and 2015 Population Statistics. (2016, April 30). Retrieved May 01, 2016, from <http://www.internetworldstats.com/africa.htm#gh>

Global forum to improve developing country access to medical devices. (2010, September 9). Retrieved May 01, 2016, from http://www.who.int/mediacentre/news/notes/2010/medical_devices_20100908/en/

Kuruk, P. (2012). Creation of 45 New Constituencies: Matters Arising (Part 3). Retrieved May 01, 2016, from <http://www.ghanaweb.com/GhanaHomePage/features/Creation-of-45-New-Constituencies-Matters-Arising-Part-3-250623>

O. (2015). HTM Engine Tutorial: Traffic Anomalies. Retrieved April 28, 2016, from https://www.youtube.com/watch?v=lzJd_a6y6-E

Appendix

This paper is dependent on HTM Algorithms that are constantly being improved by an active open-source community under the guidance of Numenta (Numenta.com, Numenta.org). The researcher therefore finds it misleading to give the steps to which MATE, as implemented in this research can be achieved. Instead, the researcher under the guidance of his supervisor is implementing MATE to be used as the Ashesi Health center. A benchmark implementation of MATE will therefore be available for any further analysis and research at the Ashesi Health center.

For further understanding of how HTM Algorithms work, the researcher finds it appropriate to direct any curious mind to numenta.org or numenta.com where up to date works on HTM are available.